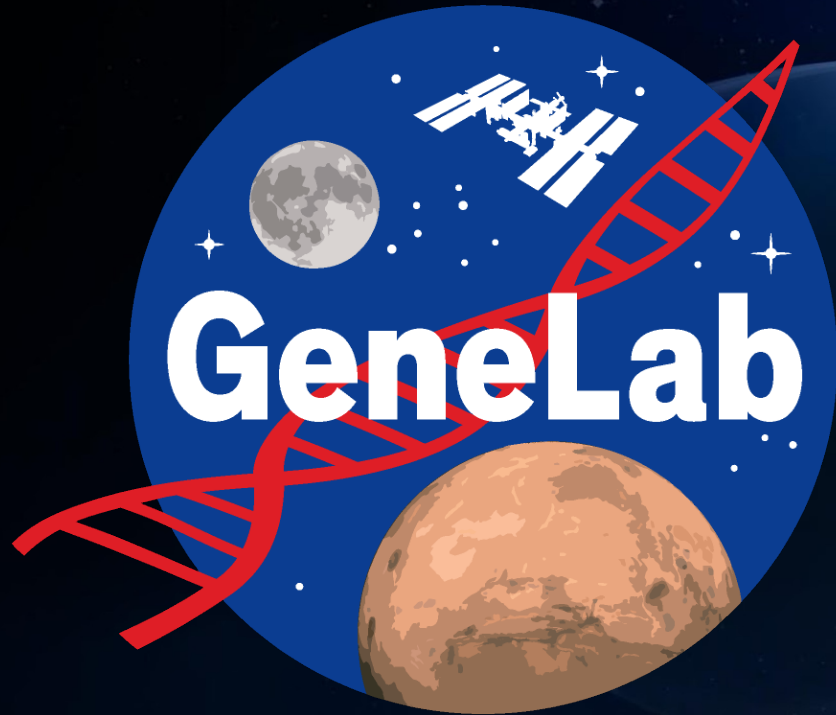


Metrics for RNA Sequencing Dataset Quality



2023 AWG Workshop – Session 1

Amanda M. Saravia-Butler, Ph.D.

Science, Data Processing, and GL4U Lead, NASA GeneLab

Contractor: KBR

National Aeronautics and
Space Administration



(bulk) RNAseq Metrics To Date



Metrics Collected and Compiled with MultiQC – Current RNAseq Pipeline (version F):

- Raw FastQC
- Trimmed FastQC
- Alignment logs
- RSEM stats
- RSeQC: geneBody coverage
- RSeQC: Infer experiment
- RSeQC: Inner distance (PE only)
- RSeQC: Read distribution

Dataset / Metric Breakdown	# Datasets Used Out of Total
PE datasets with all metrics	16/18
SE datasets with all metrics	17/18
Datasets with all but RSeQC	54/56
Datasets with raw, trimmed, aligned only	1/1
Datasets with raw only	18/19

Raw / Trimmed Metrics



Raw/Trimmed FastQC Metrics Calculated From the MultiQC Report Data (and HTStream):

Total Studies: 106 (Raw), 88 (Trimmed)

➤ Total Seqs (Read Depth)

➤ % rRNA Contamination (HTStream Analysis)

➤ Mean Read Length

➤ Average Q Score (Across all plotted base positions)

➤ Median Q Score (Across all plotted base positions)

➤ % Duplicates

➤ % N Content Summed Across All Plotted Bases Positions, Normalized for Read Length

➤ Mean %GC

➤ Min %GC reaching 1% Counts Raw

➤ Max %GC reaching 1% Counts Raw

➤ 25% Quartile AUC Point %GC Raw

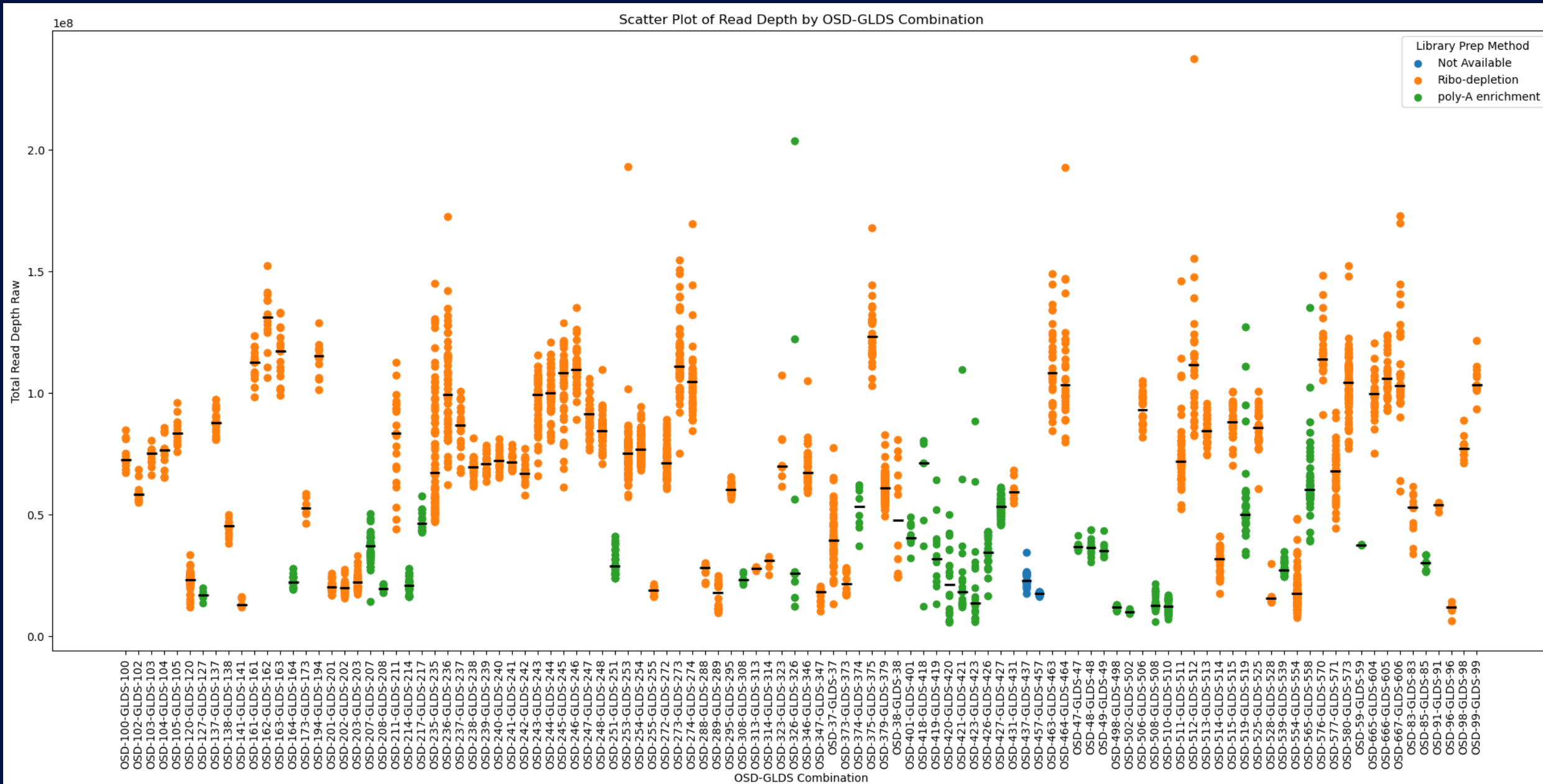
➤ 50% Quartile AUC Point %GC Raw

➤ 75% Quartile AUC Point %GC Raw

*Data plotted with samples grouped based on OSD/GLDS, organism, organism/tissue; samples colored based on library preparation method

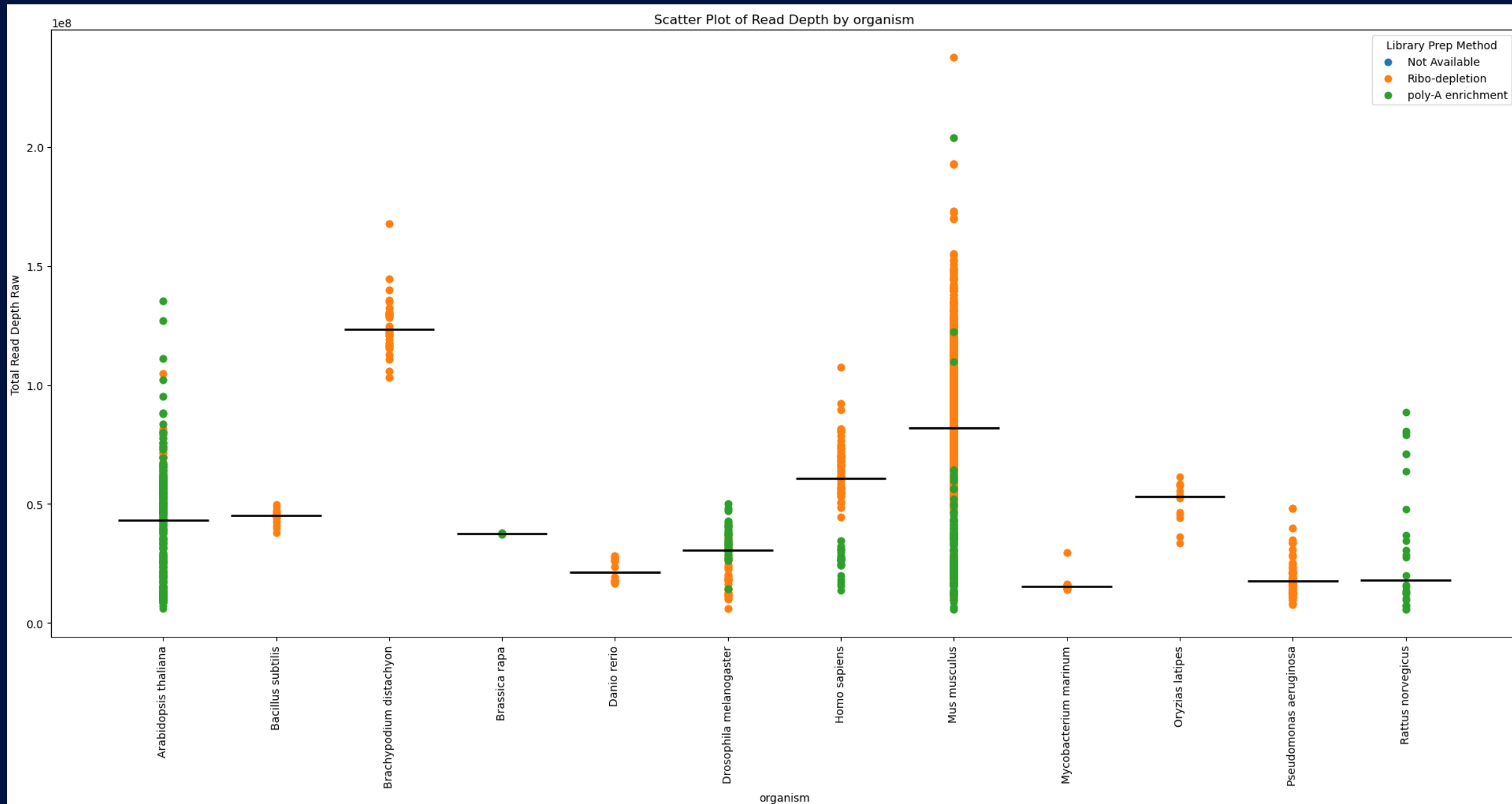


Raw Read Depth – OSD/GLDS



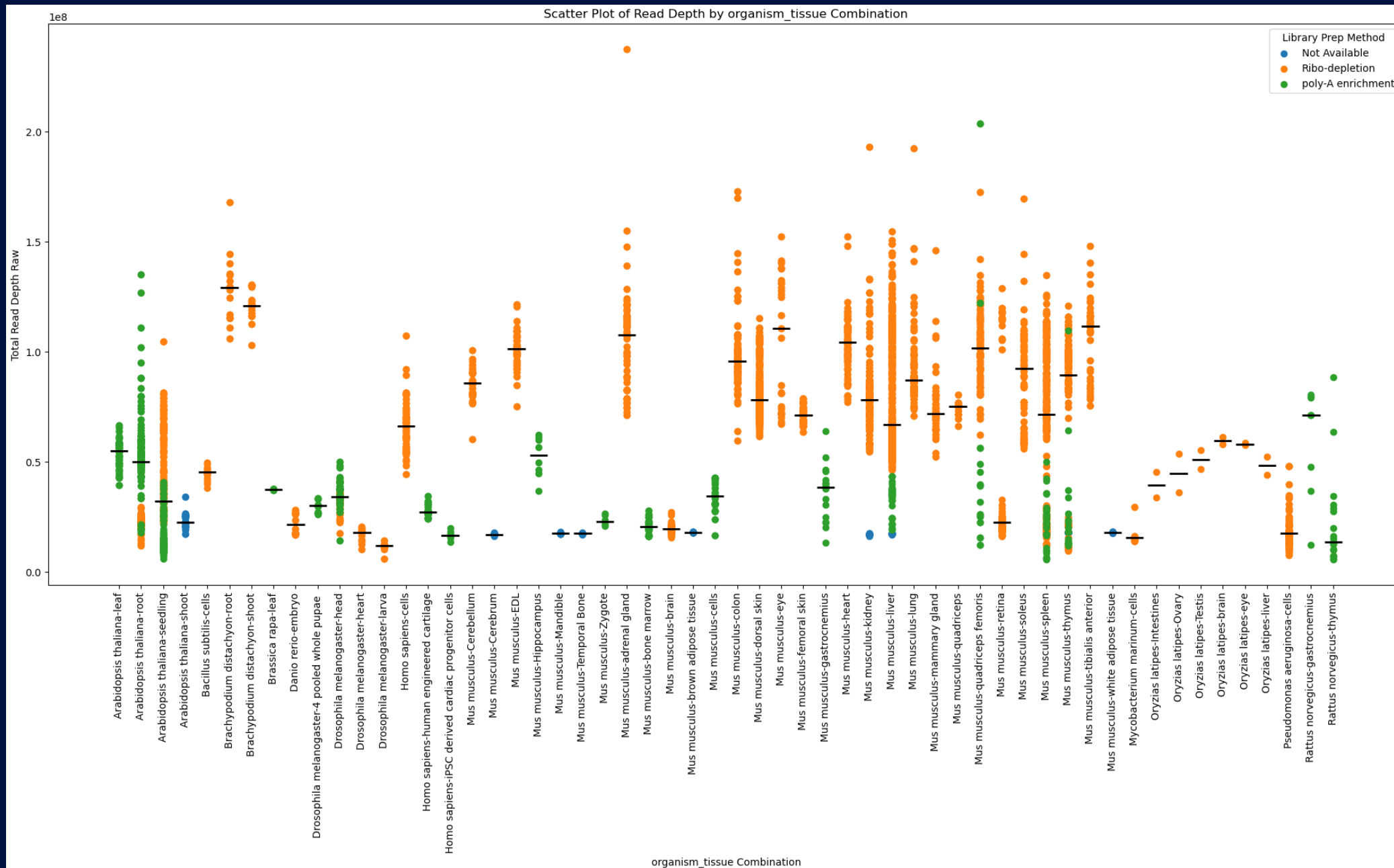


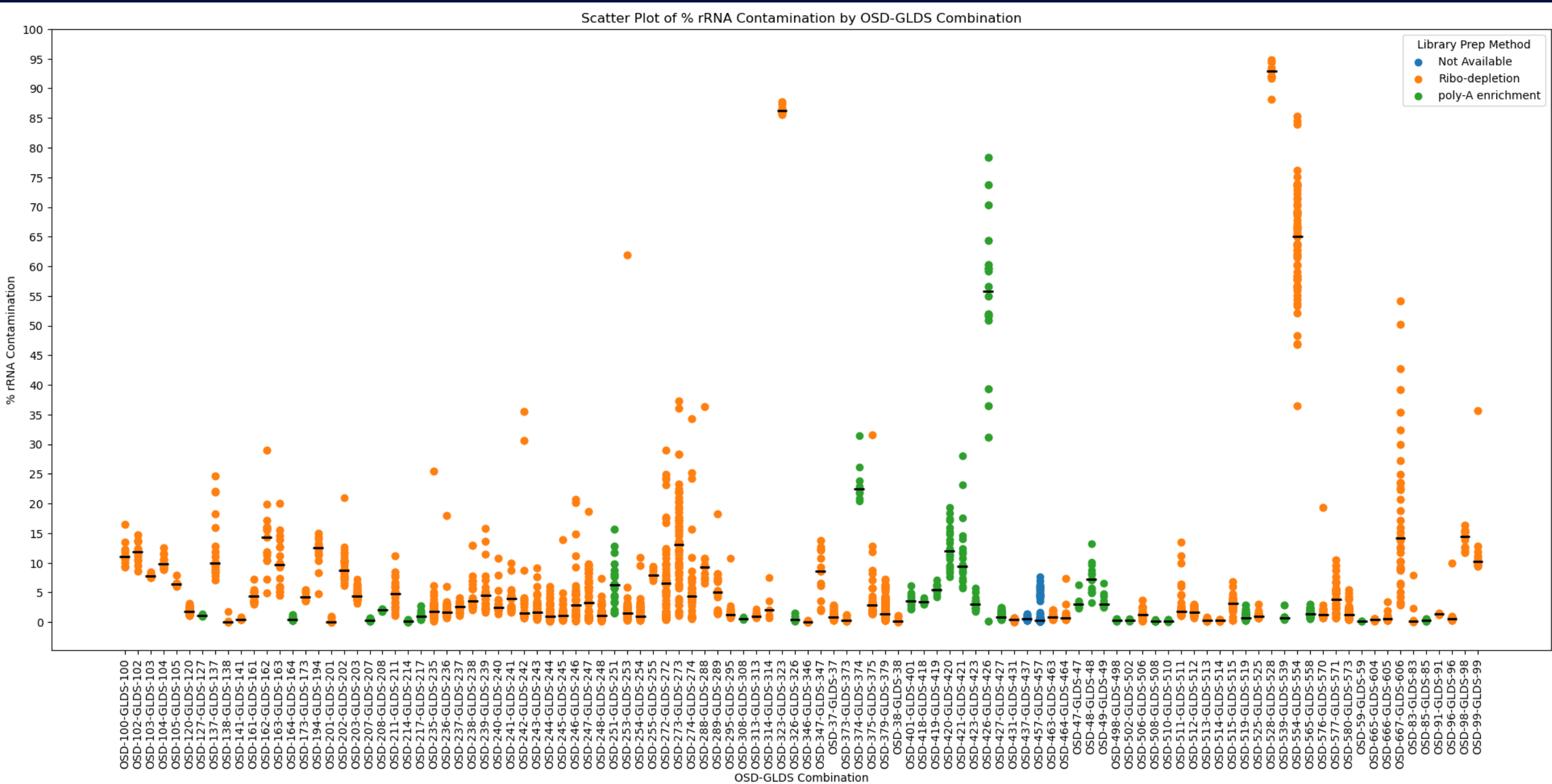
Raw Read Depth – Organism





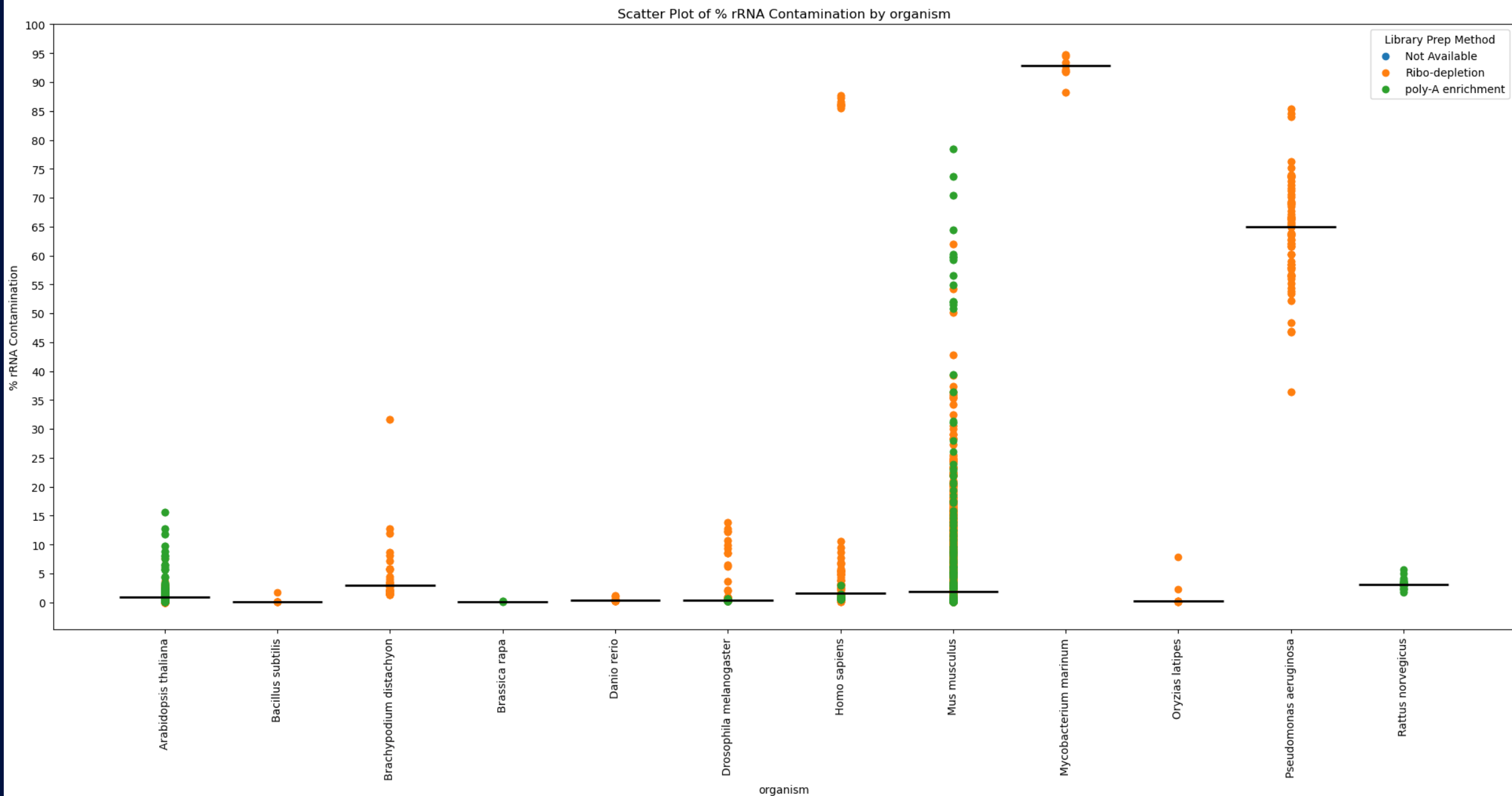
Raw Read Depth – Organism/Tissue



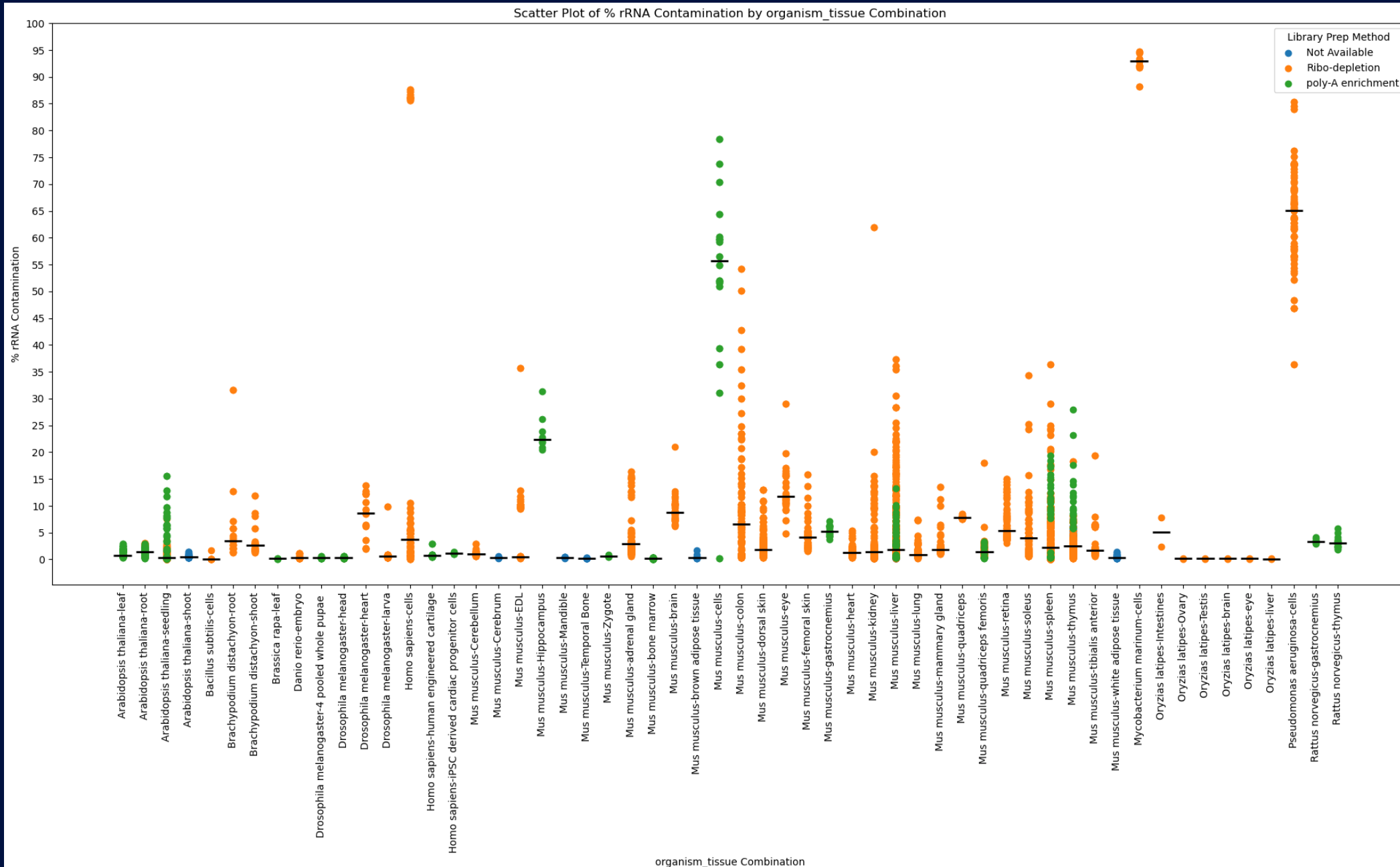




Raw %rRNA Contamination – Organism



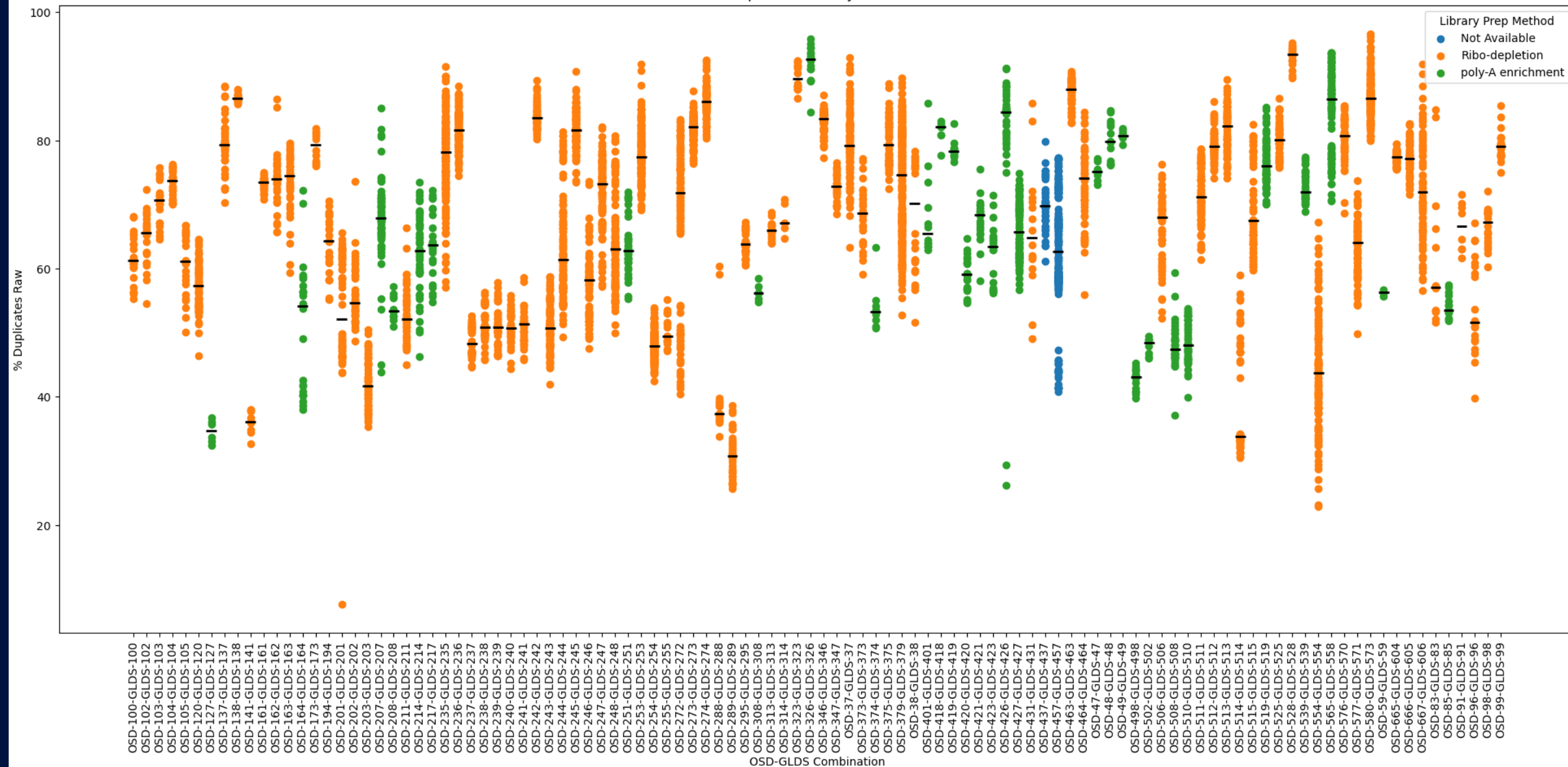
Raw %rRNA Contamination – Organism/Tissue





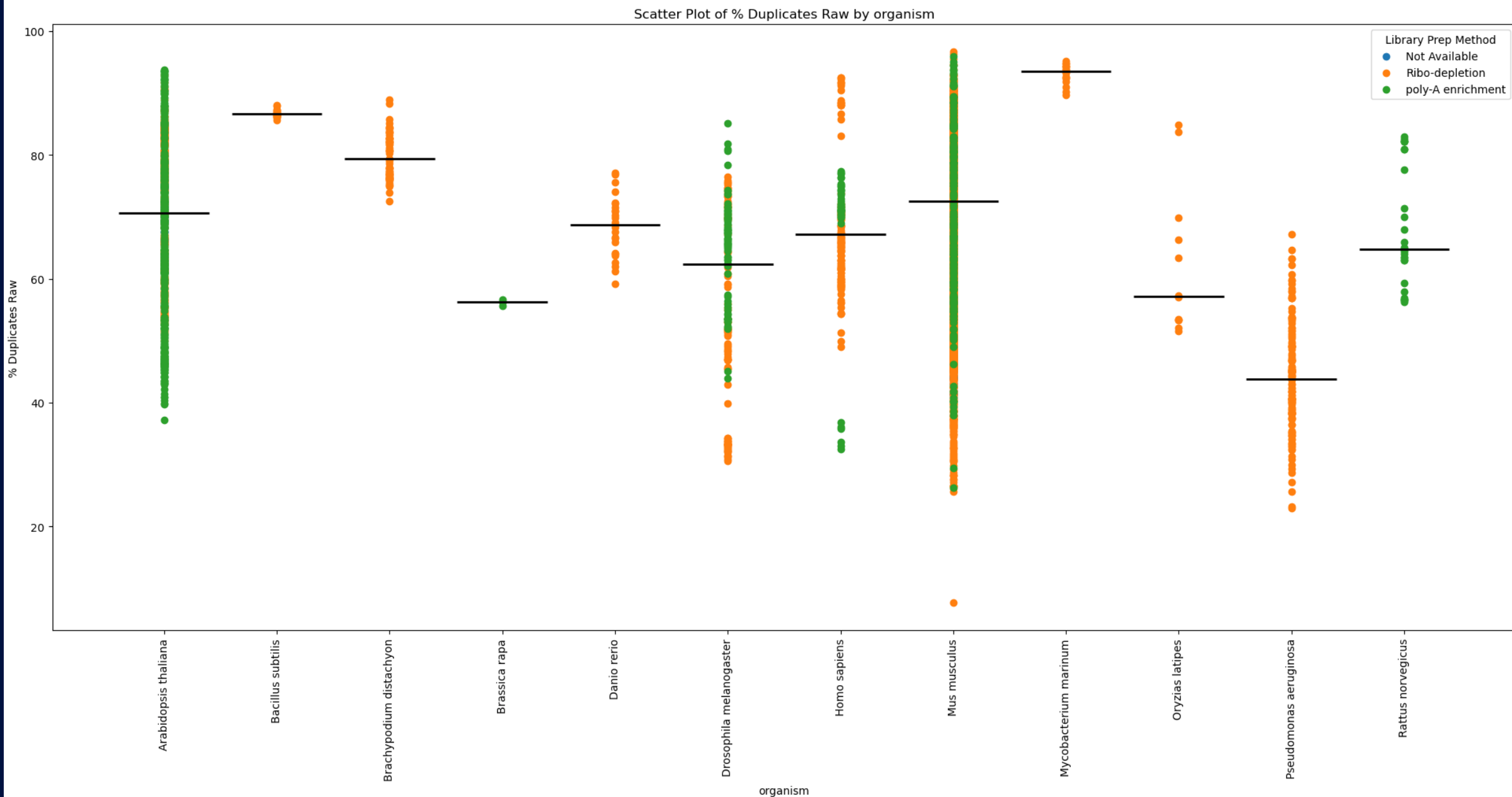
Raw % Duplicates – OSD/GLDS

Scatter Plot of % Duplicates Raw by OSD-GLDS Combination



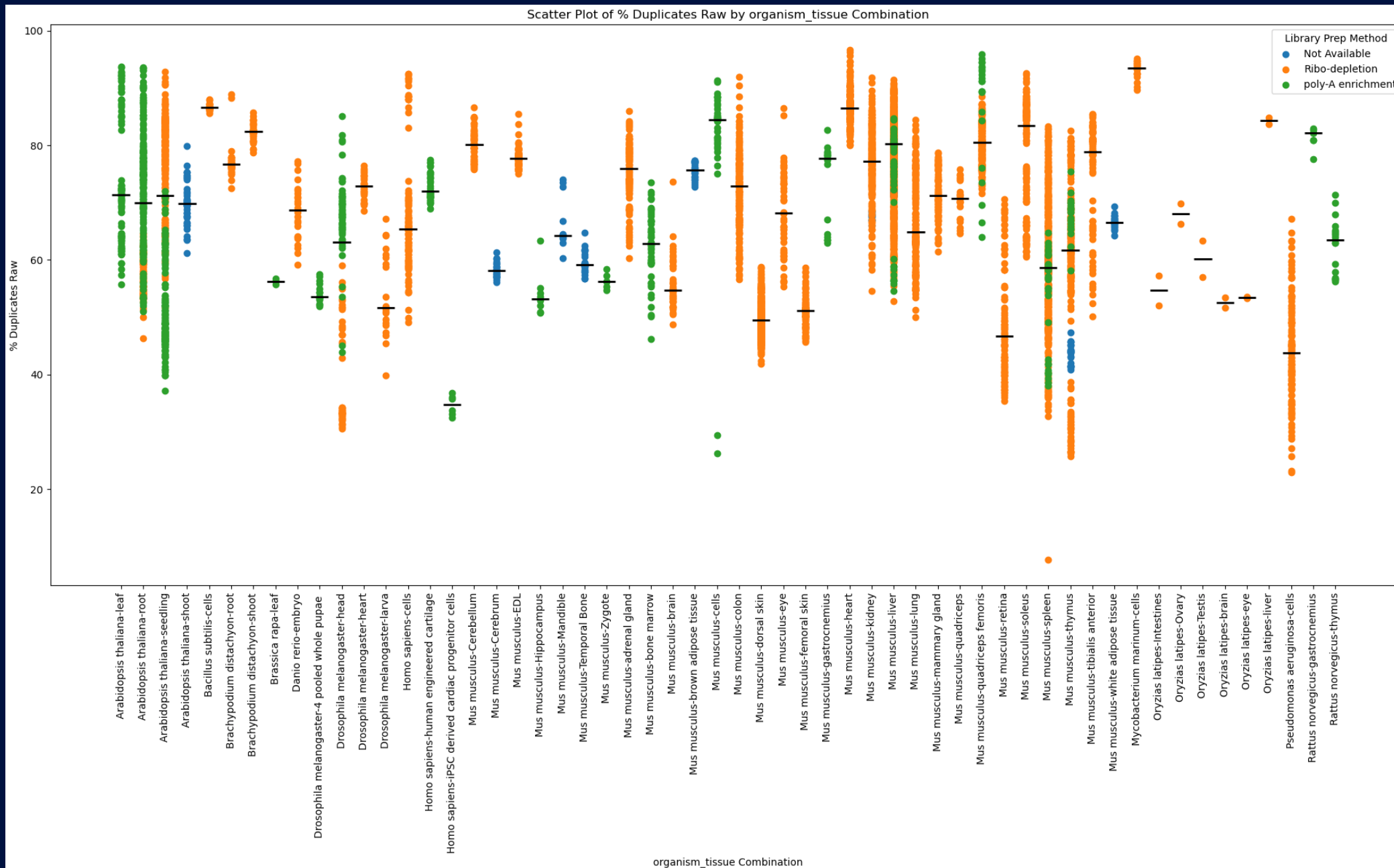


Raw % Duplicates – Organism





Raw % Duplicates – Organism/Tissue



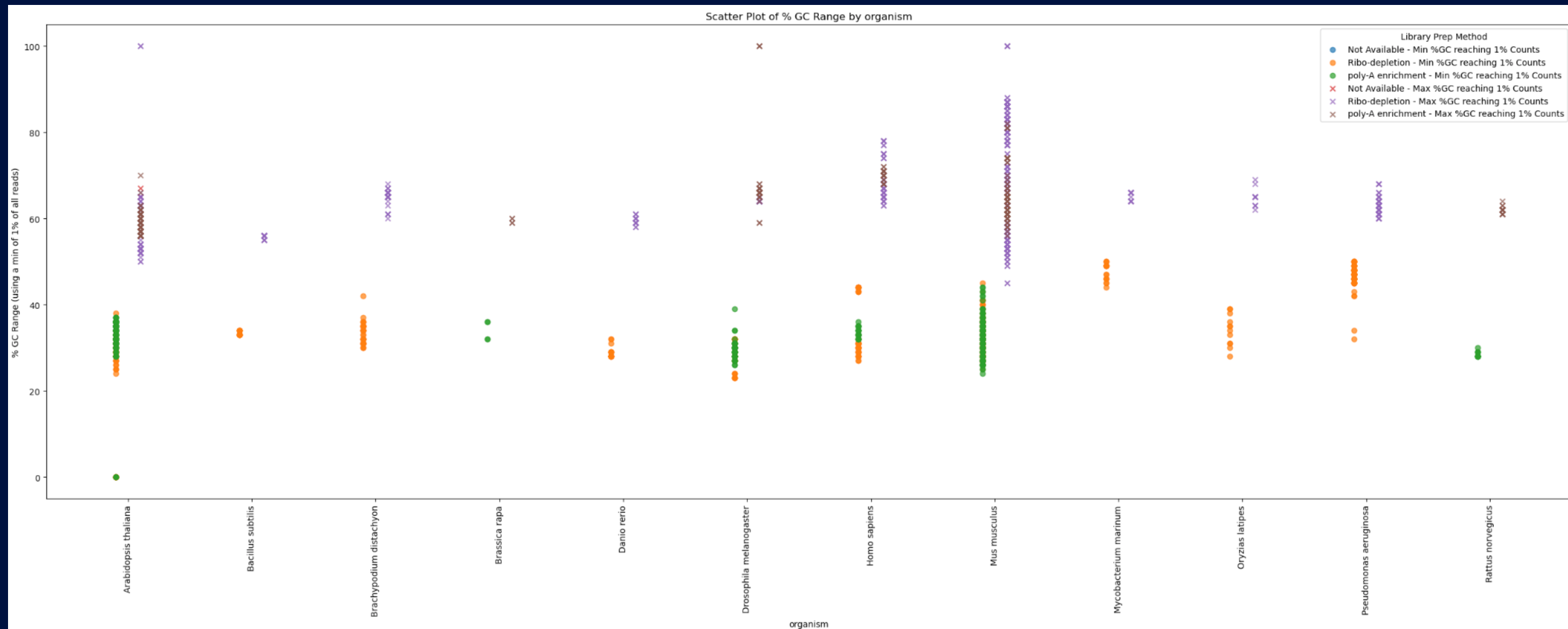


Raw % GC Range – OSD/GLDS



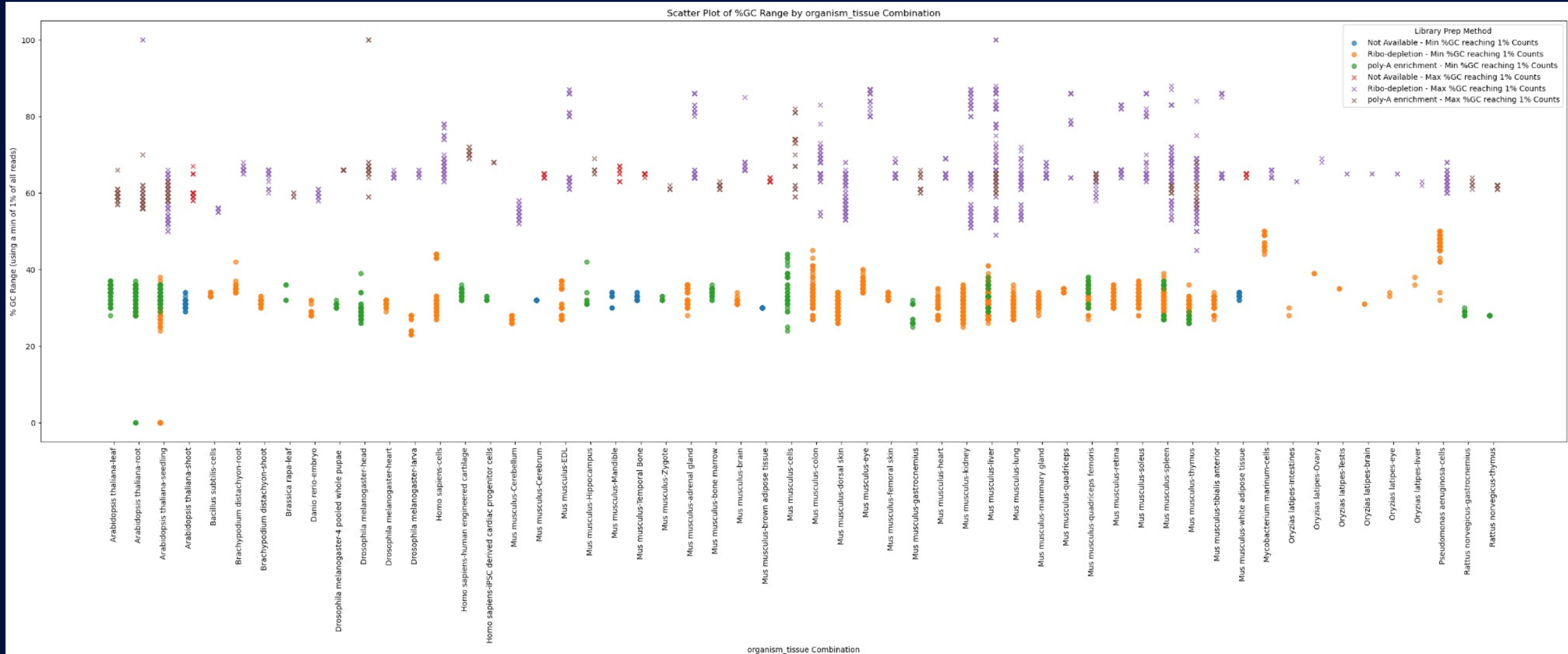


Raw % GC Range – Organism





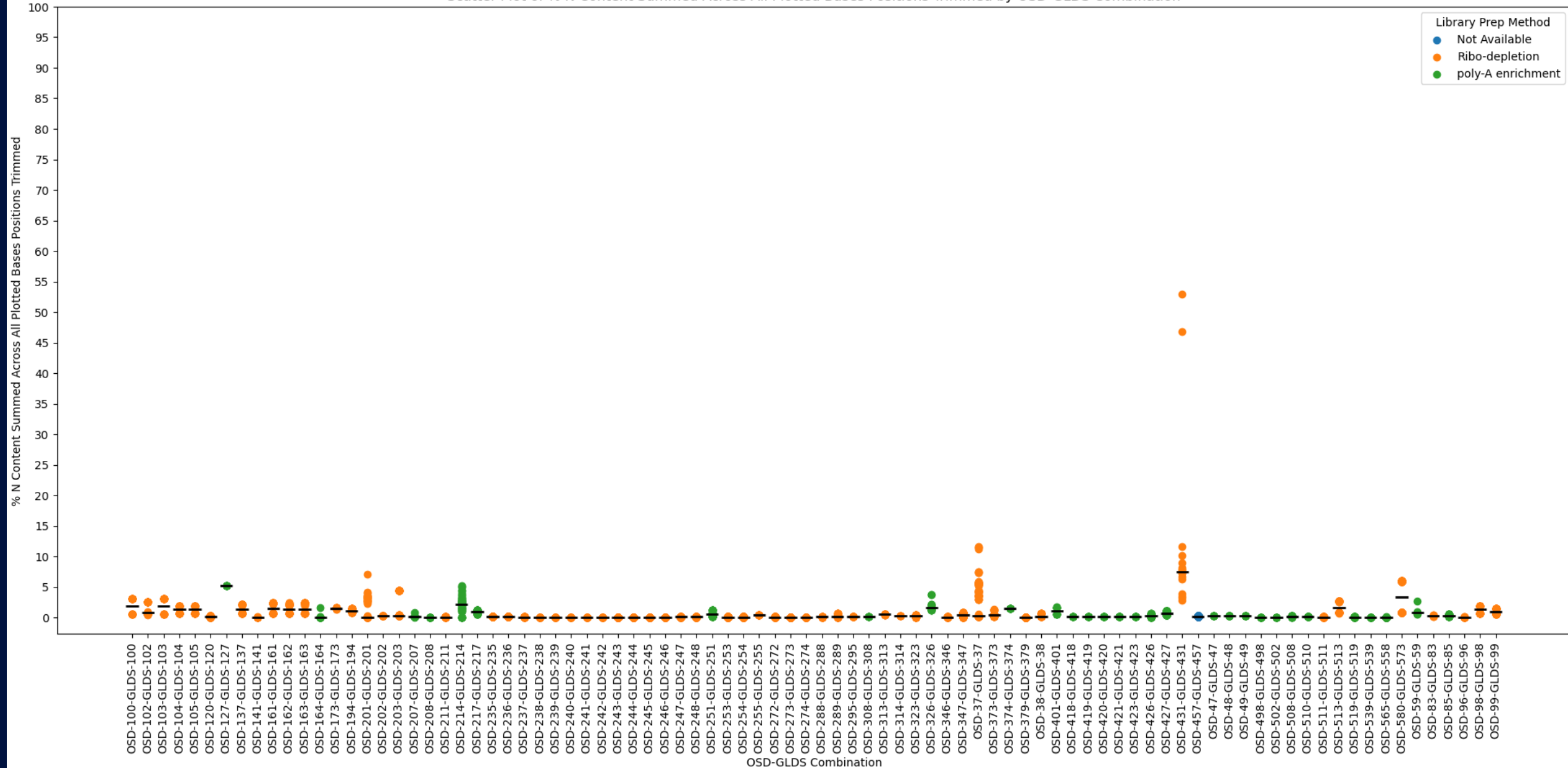
Raw % GC Range – Organism/Tissue





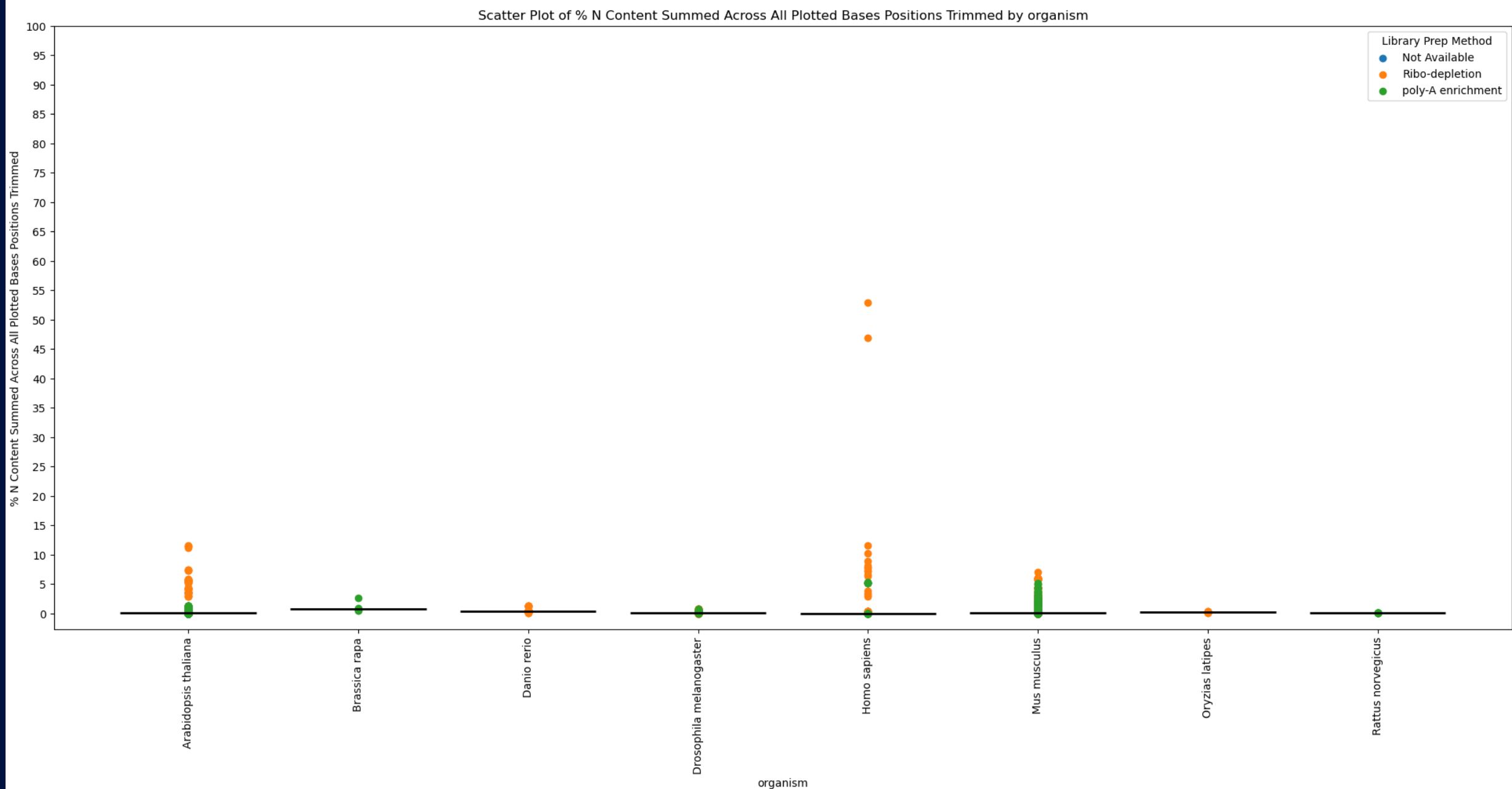
Trimmed %N Content – OSD/GLDS

Scatter Plot of % N Content Summed Across All Plotted Bases Positions Trimmed by OSD-GLDS Combination



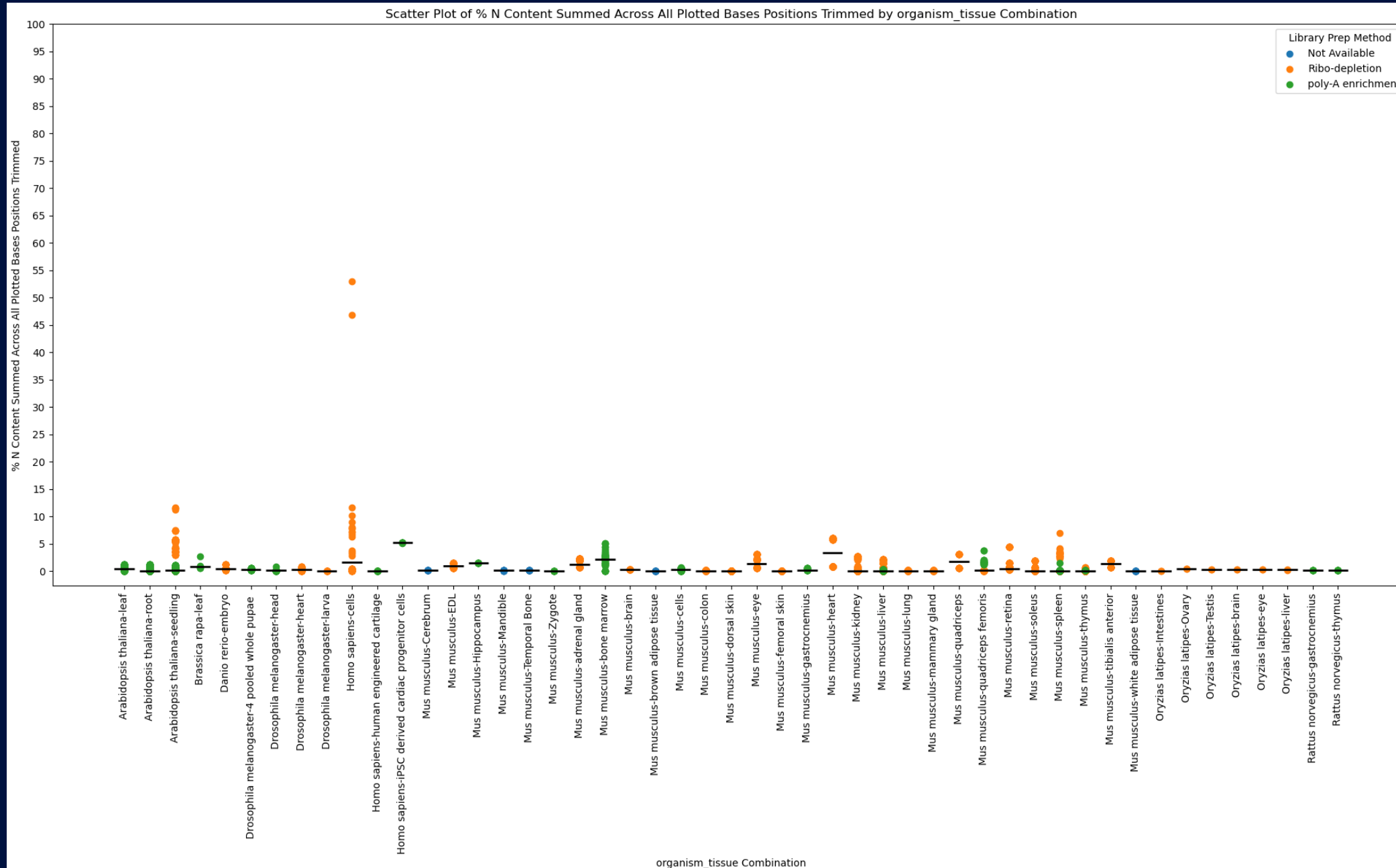


Trimmed %N Content – Organism





Trimmed %N Content– Organism/Tissue



Alignment Metrics

Aligned Data QC Metrics Collected



Alignment Metrics Calculated From the MultiQC Report Data of Alignment Logs:

88 Total Studies

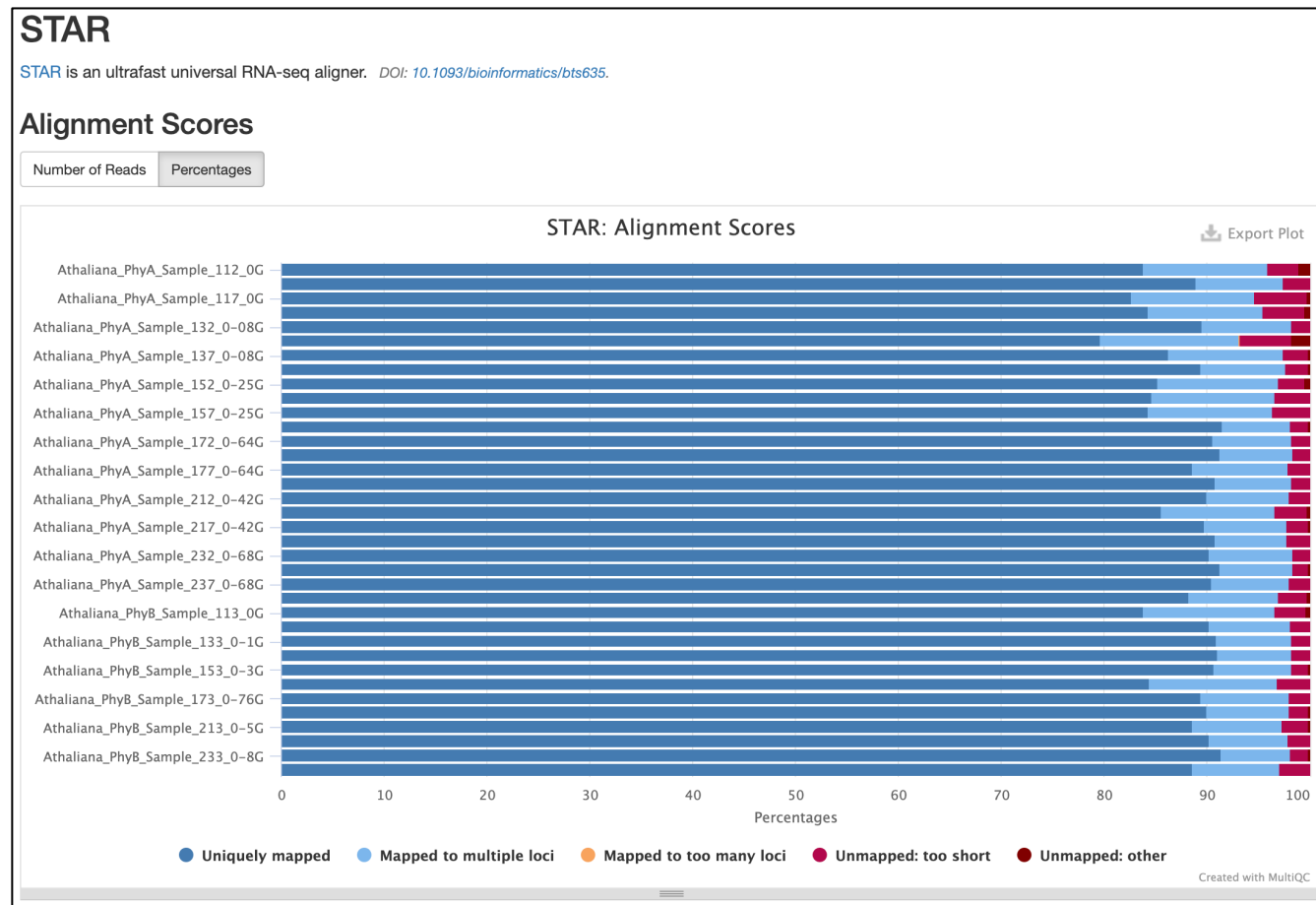
➤ % Uniquely mapped

➤ % Mapped to multiple loci

➤ % Mapped to too many loci

➤ % Unmapped too short

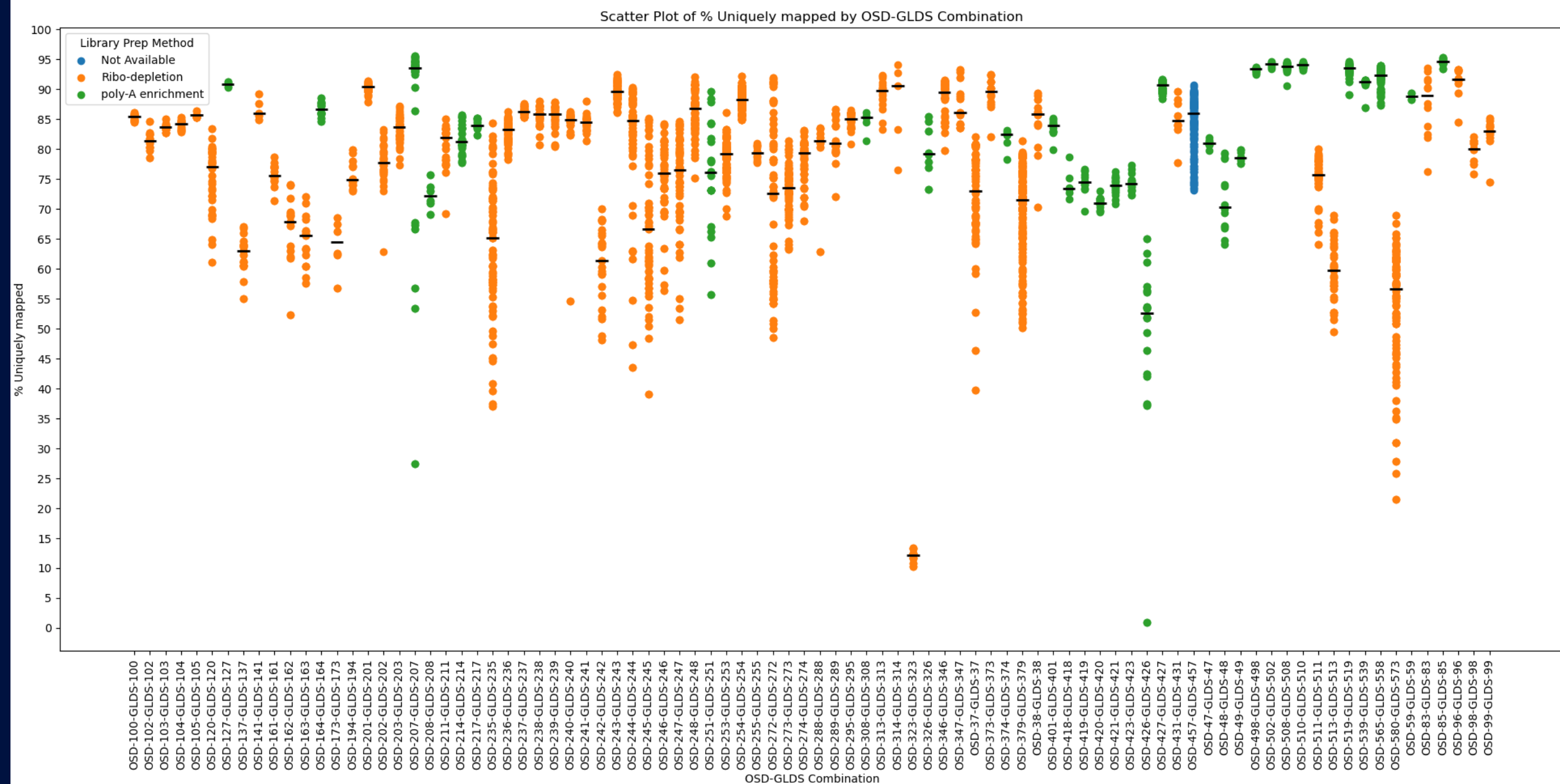
➤ % Unmapped other



*Data plotted with samples grouped based on OSD/GLDS, organism, organism/tissue; samples colored based on library preparation method

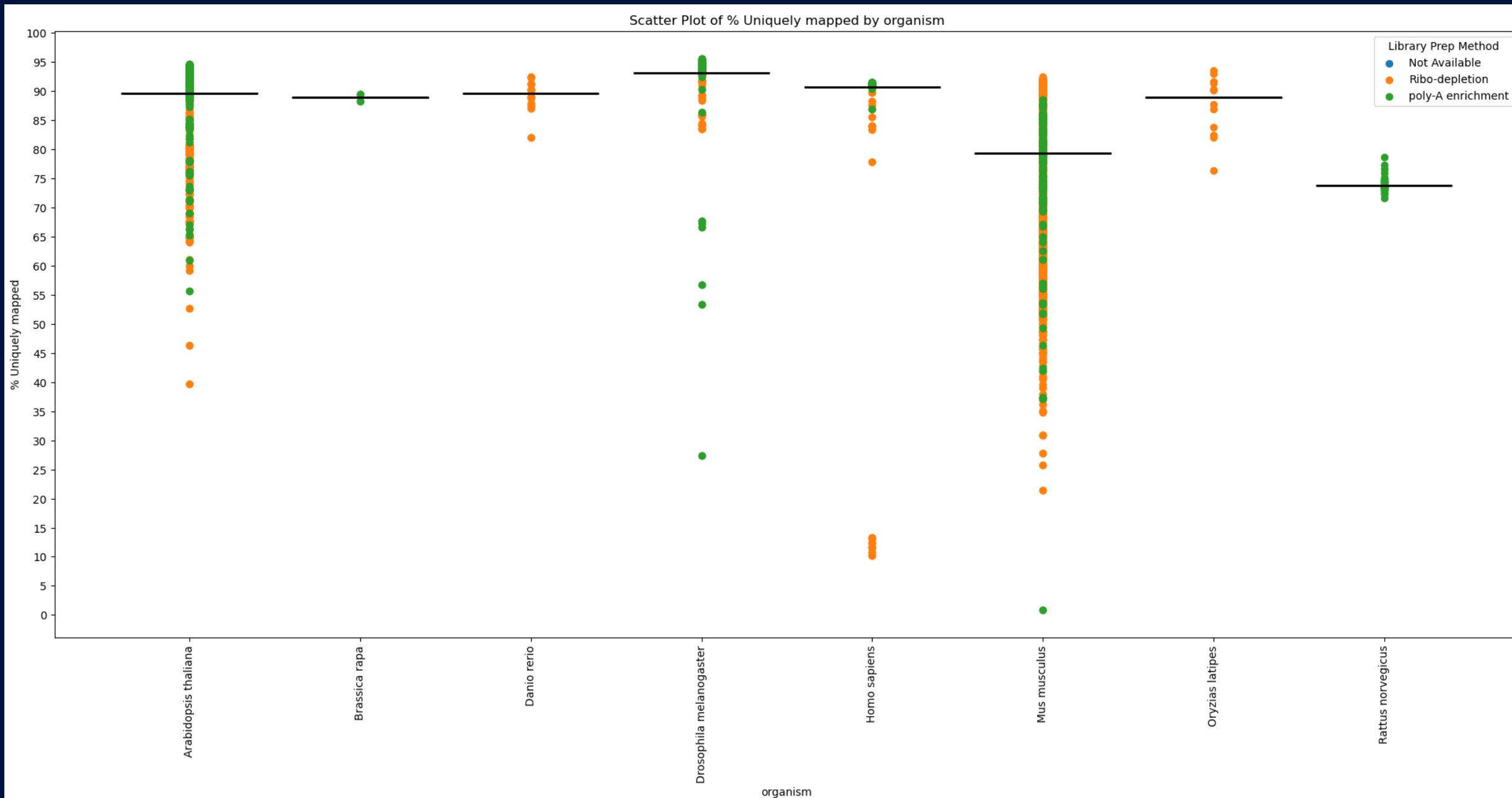


% Uniquely Mapped Reads – OSD/GLDS



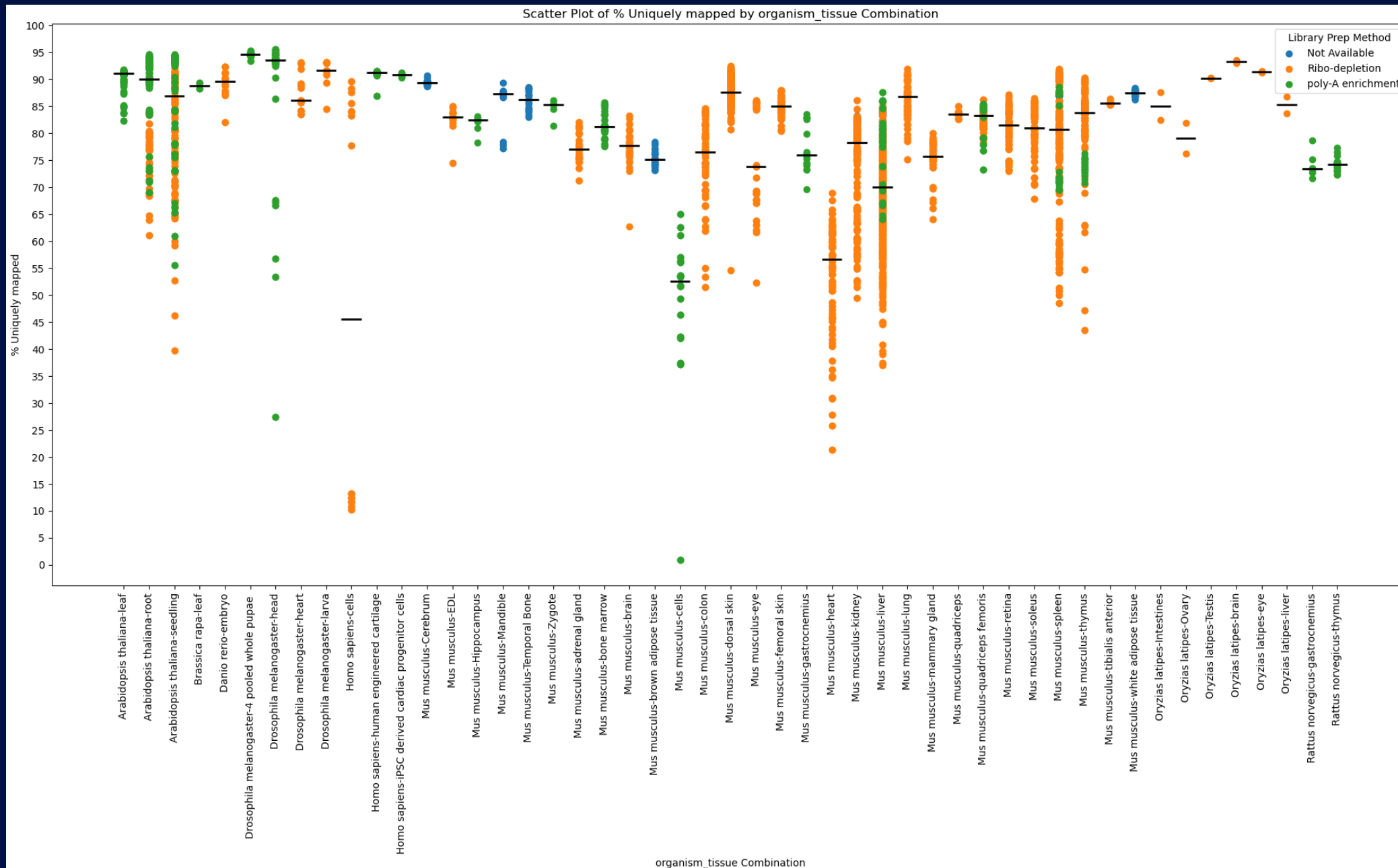


% Uniquely Mapped Reads – Organism





% Uniquely Mapped Reads – Organism/Tissue



RSEM Stats Metrics

RSEM Quantitation QC Metrics Collected



RSEM Count Metrics Calculated From the MultiQC Report Data of RSEM stats:

87 Total Studies

➤ % Aligned uniquely to a gene

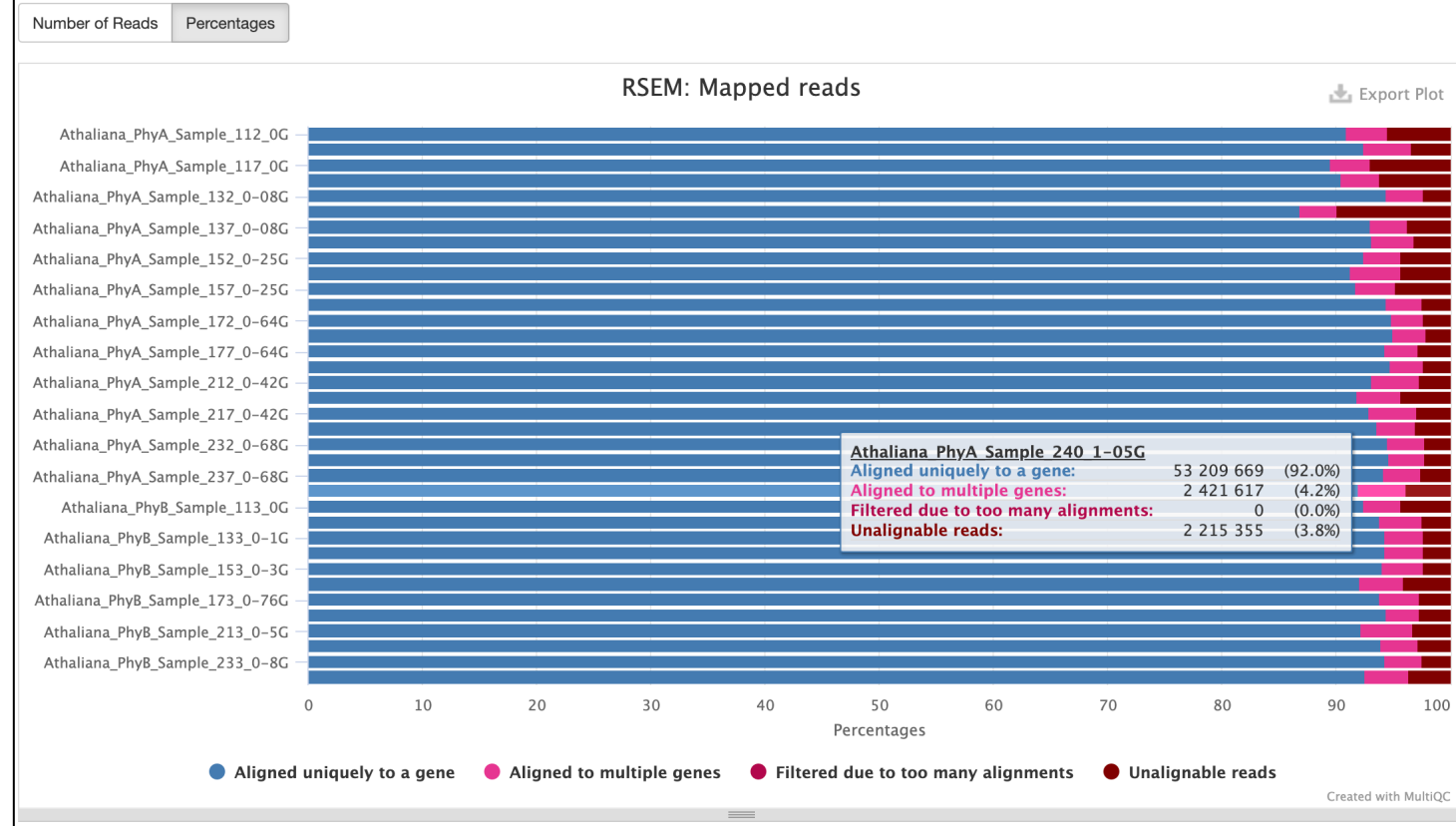
➤ % Aligned to multiple genes

➤ % Filtered due to too many alignments

➤ % Unalignable reads

Mapped Reads

A breakdown of how all reads were aligned for each sample.

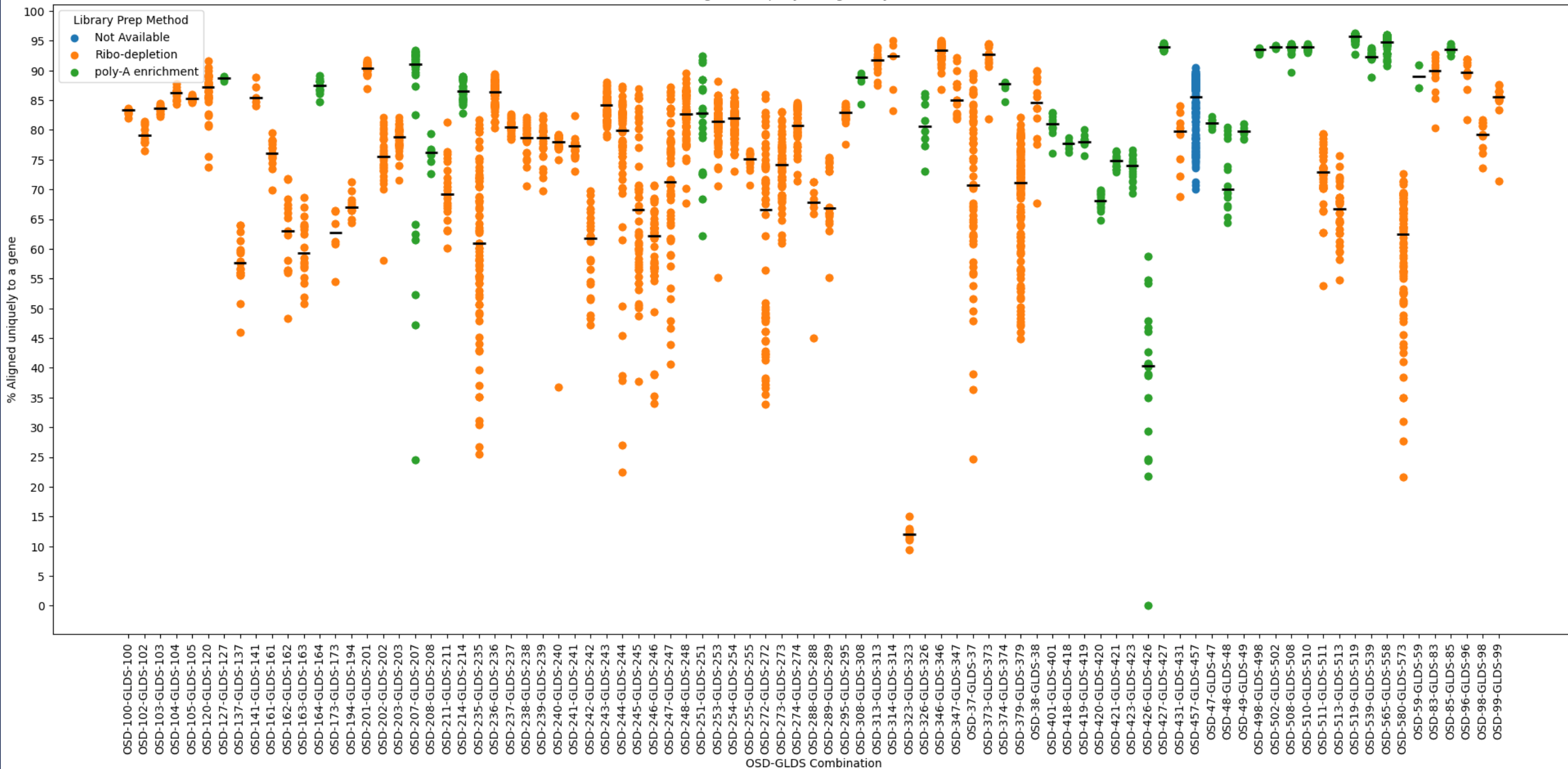


*Data plotted with samples grouped based on OSD/GLDS, organism, organism/tissue; samples colored based on library preparation method



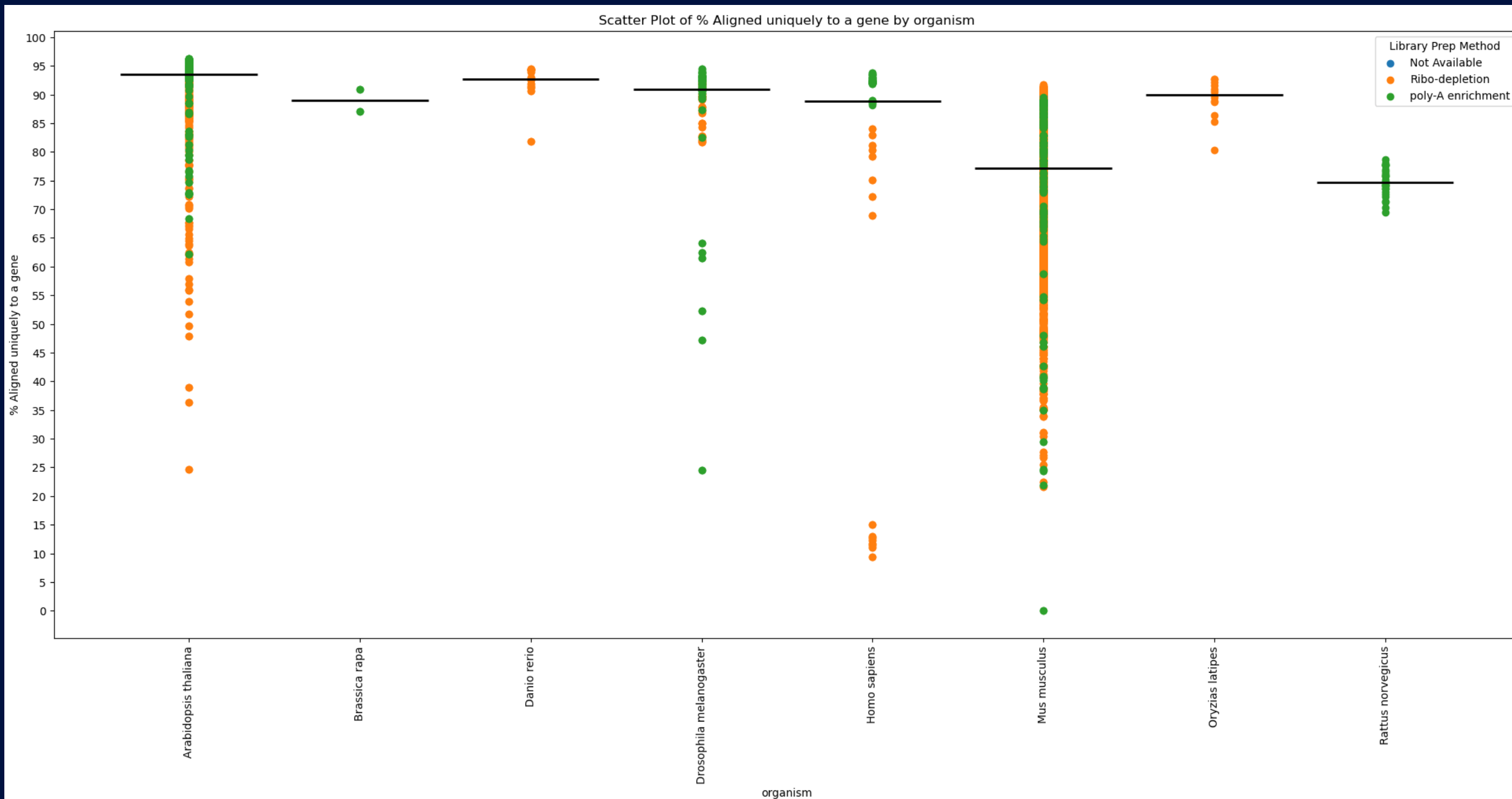
% Reads Aligned Uniquely To A Gene – OSD/GLDS

Scatter Plot of % Aligned uniquely to a gene by OSD-GLDS Combination

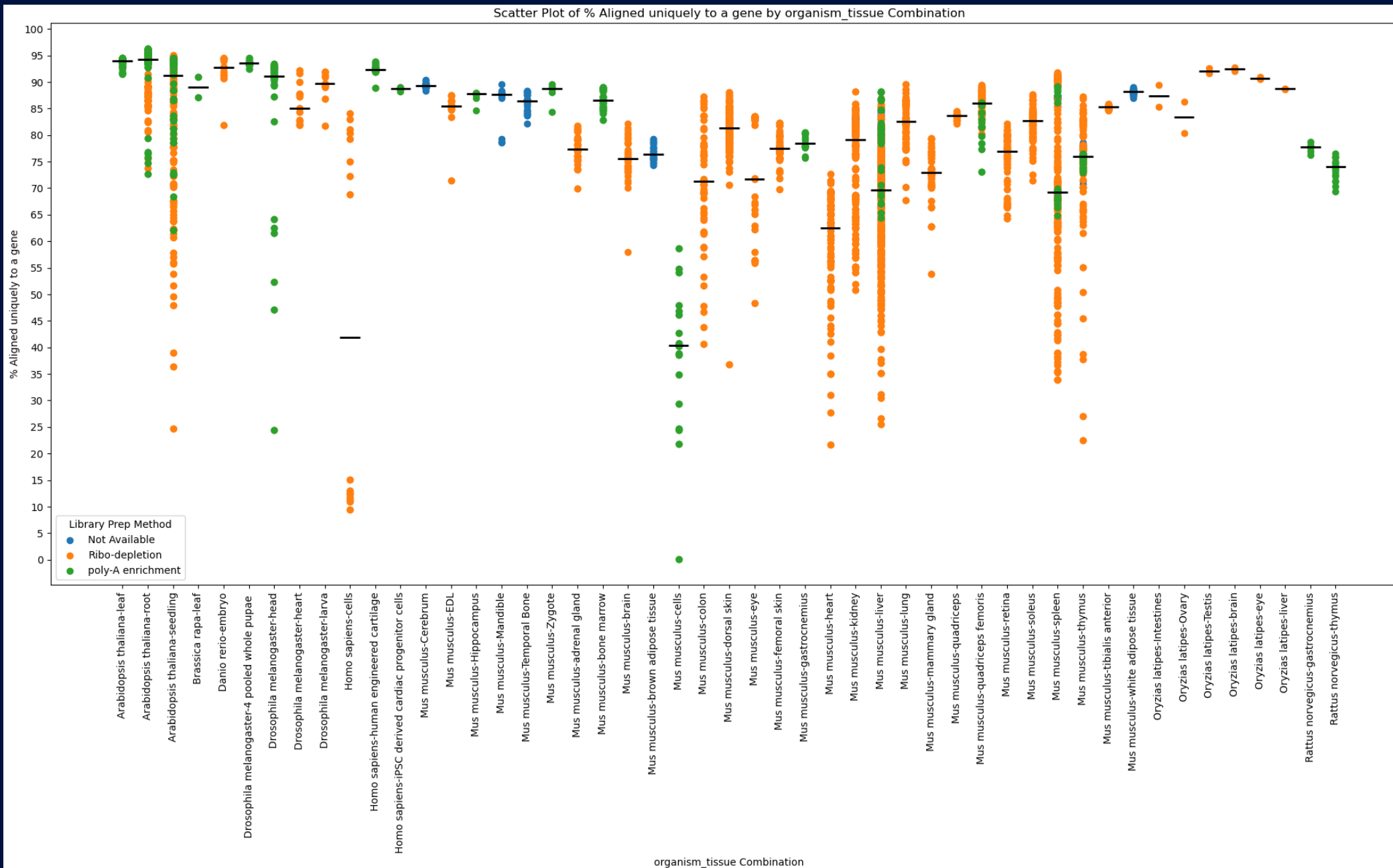




% Reads Aligned Uniquely To A Gene – Organism



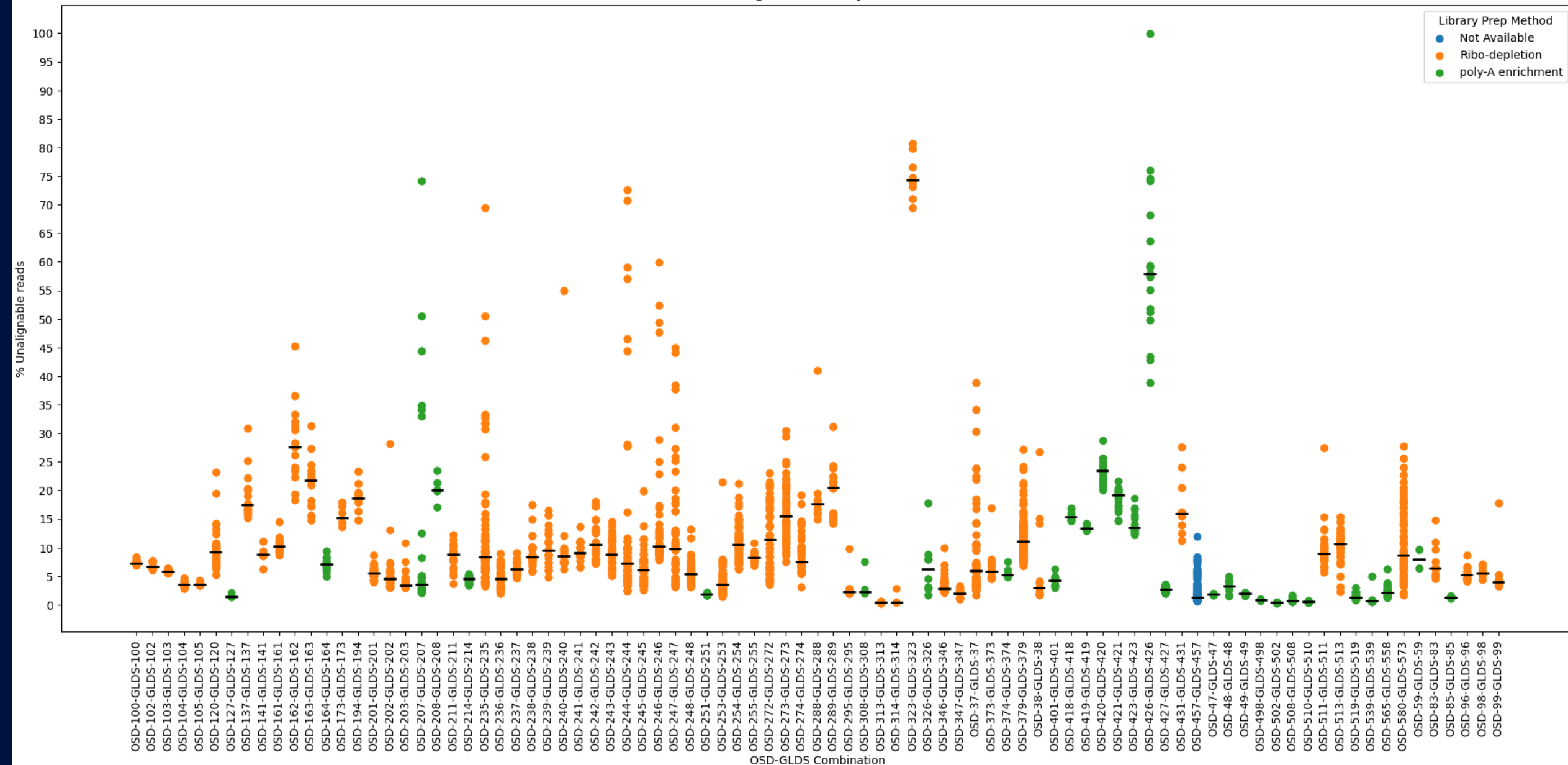
% Reads Aligned Uniquely To A Gene – Organism/Tissue





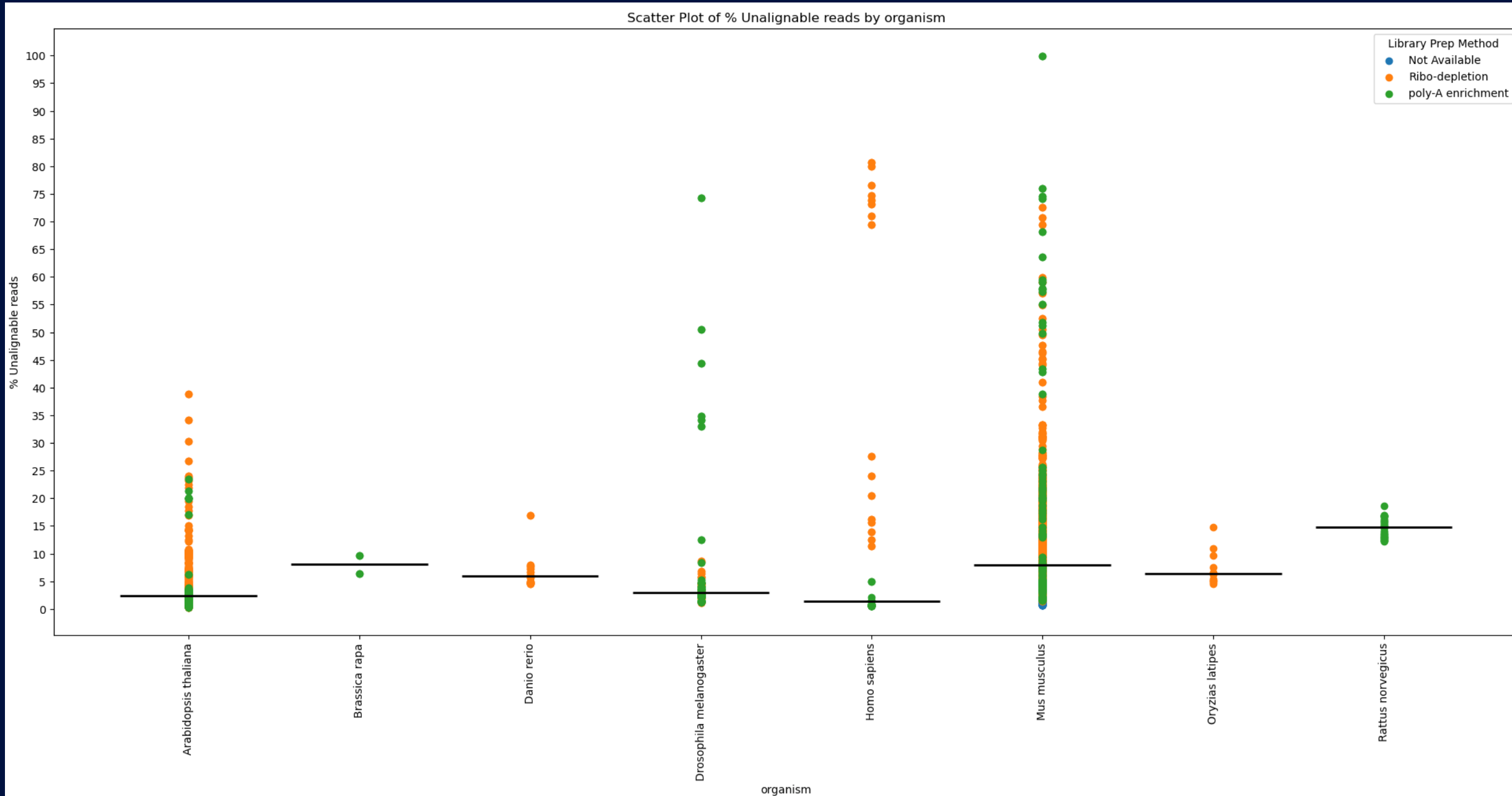
% Unalignable Reads – OSD/GLDS

Scatter Plot of % Unalignable reads by OSD-GLDS Combination



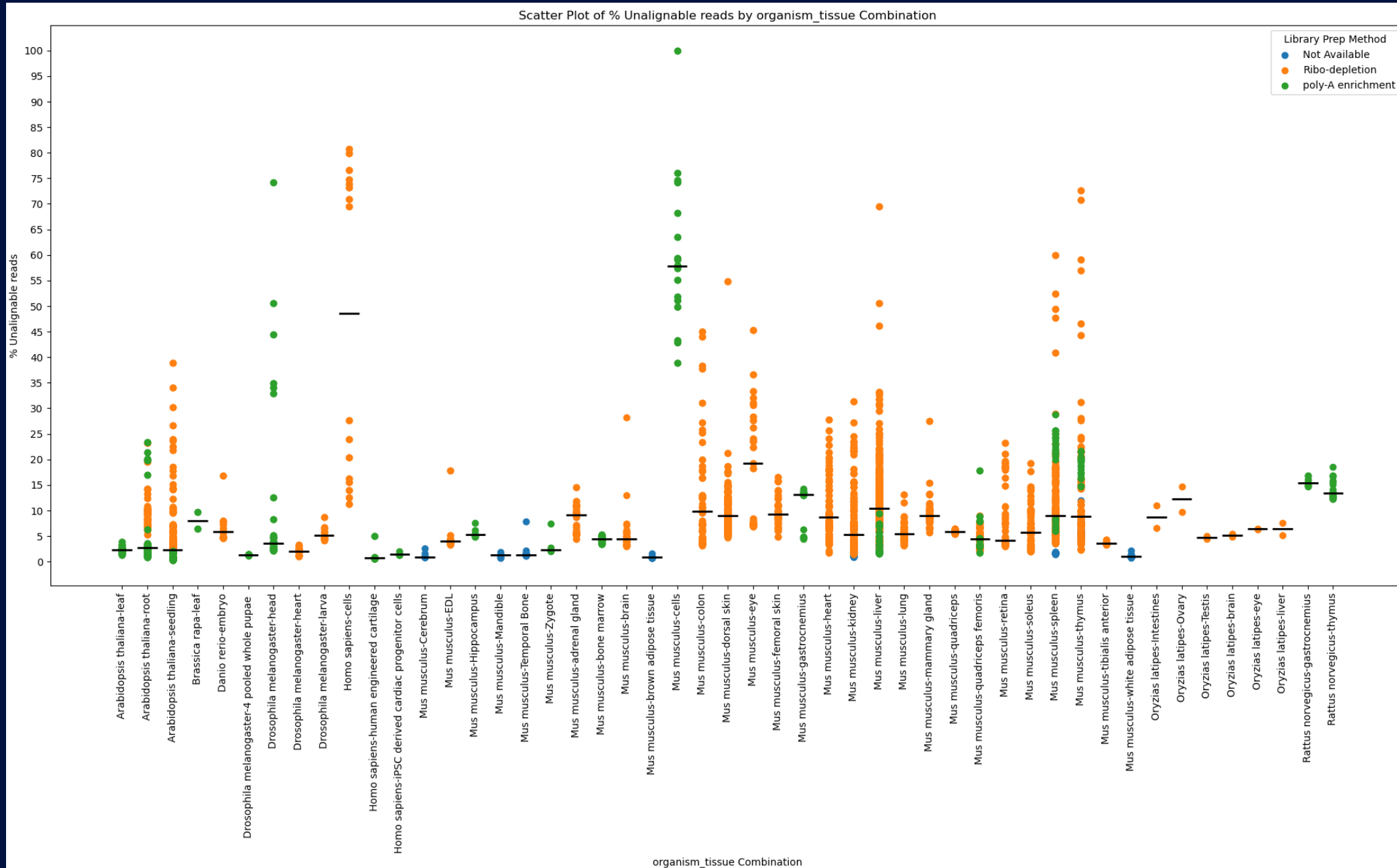


% Unalignable Reads – Organism





% Unalignable Reads – Organism/Tissue



RNAseq Metrics Breakout Session

RNAseq Metrics Breakout Session Tasks



- Review and provide feedback for the proposed RNAseq pipeline updates
- For each set of proposed (sample-level) metrics listed below, answer the following questions:
 - Should we keep all metrics listed?
 - Should we add any metrics?
 - Are the values defined for green, yellow, and red flags appropriate?
- How should these metrics be displayed on OSDR at both the sample and dataset levels?
 - Sample level: Should we have an overall quality score for each sample or should we have a grid showing the quality score for each metric?
 - Dataset level: How should the sample-level data be compiled to generate a dataset-level score? What other metrics should be considered at the dataset level (e.g. spread of data across samples)?

	Read Depth	%rRNA Contamination	Median Q score	% Duplications	% GC Range	% N Content	% Uniquely Mapped	% Aligned Uniquely	% Unalignable
Sample 1									
Sample 2									
Sample 3									

MultiQC Data Source	Metric	Green-Flag	Yellow-Flag	Red-Flag
Raw	M Seqs (read depth)	ribo-deplete: <i>Mammals</i> : 54-66M (60M +/- 10%) <i>Non-mammals</i> : 36-44M (40M +/- 10%) <i>Plants</i> : 54-66M (60M +/- 10%) <i>Bacteria</i> : 27-33M (30M +/- 10%) poly-A selection: <i>Mammals</i> : 36-44M (40M +/- 10%) <i>Non-mammals</i> : 27-33M (30M +/- 10%) <i>Plants</i> : 36-44M (40M +/- 10%)	ribo-deplete: <i>Mammals</i> : 44-53.9M (49M +/- 10%) <i>Non-mammals</i> : 29-35.9M (32.5M +/- 10%) <i>Plants</i> : 44-53.9M (49M +/- 10%) <i>Bacteria</i> : 21-26.9M (24M +/- 10%) poly-A selection: <i>Mammals</i> : 28-35.9M (32M +/- 10%) <i>Non-mammals</i> : 21-26.9M (24M +/- 10%) <i>Plants</i> : 28-35.9M (32M +/- 10%)	ribo-deplete: <i>Mammals</i> : <44M <i>Non-mammals</i> : <29M <i>Plants</i> : <44M <i>Bacteria</i> : <21M poly-A selection: <i>Mammals</i> : <28M <i>Non-mammals</i> : <21M <i>Plants</i> : <28M
Raw	%rRNA contamination	0-10%	11-30%	31-100%
Raw	Median Q score	>30	26-30	<26
Raw	% Dups	<50%	50-70%	>70%
Raw	% GC range (using a min of 1% of all reads)	30-60%	61-75%	>75%
Trimmed	% N Content (Summed across all plotted base pairs positions and normalized over total read length)	0-5%	6-15%	>15%
Aligned	% Uniquely mapped	75-100%	60-74%	<60%
RSEM Count	% Aligned uniquely to a gene	75-100%	60-74%	<60%
RSEM Count	% Unalignable reads	0-10%	11-25%	>25%

Acknowledgements

- **NASA GeneLab**
- **GeneLab Data Processing Team
(both past and present)**
- **AWGs**
- **Funding:** NASA Space Biology program within the NASA Science Mission Directorate's (SMD) Biological and Physical Sciences (BPS) Division